CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

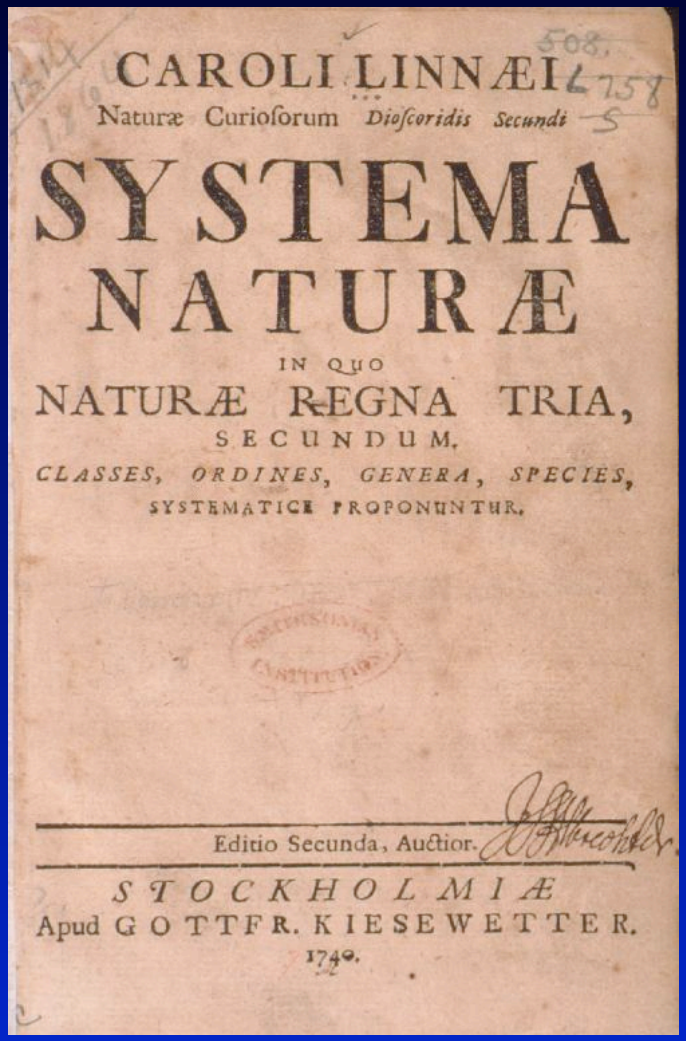
Introduction:

Theory of Evolution
Bioinformatics

Anders Gorm Pedersen
Molecular Evolution Group

Classification: Linnaeus

Carl Linnaeus
1707-1778



CAROLI LINNÆI
Naturæ Curioſorum *Dioſcoridis Secundi*

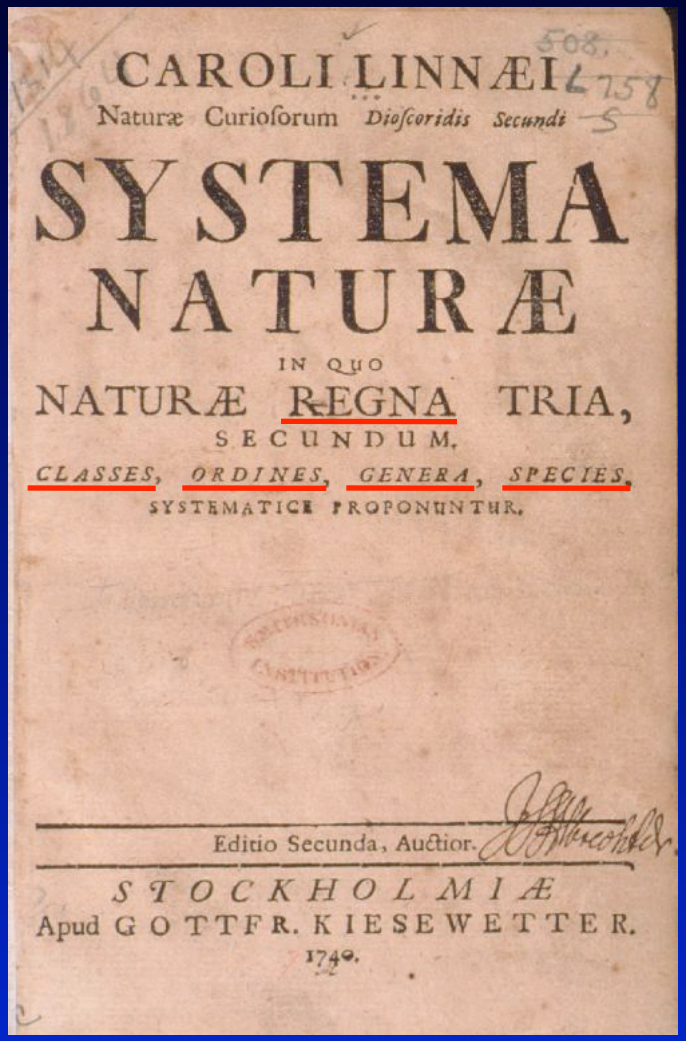
SYSTEMA NATURÆ

IN QUO
NATURÆ REGNA TRIA,
SECUNDUM,
CLASSES, ORDINES, GENERA, SPECIES,
SYSTEMATICE PROPONUNTUR.

Editio Secunda, Auctior.

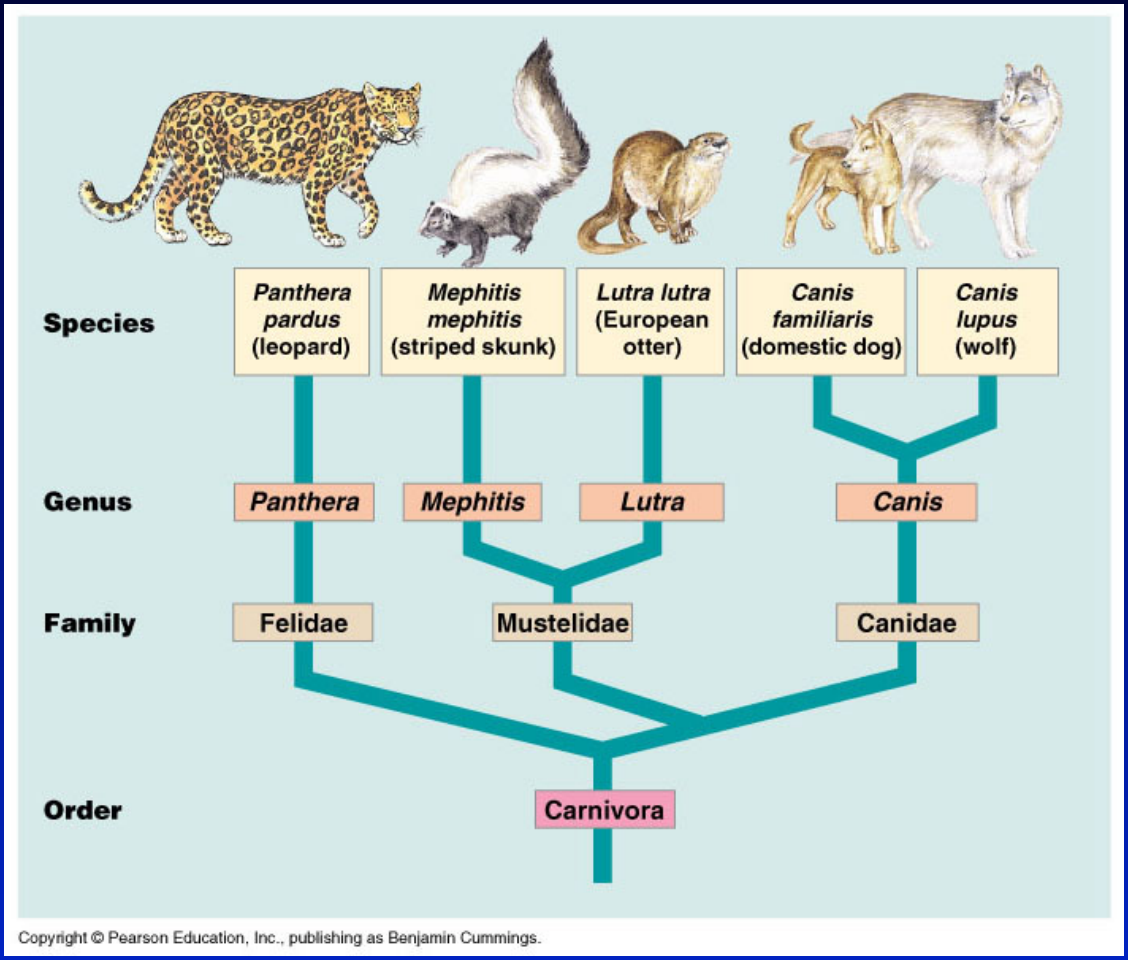
STOCKHOLMIÆ
Apud GOTTFR. KIESEWETTER.
1740.

Classification: Linnaeus

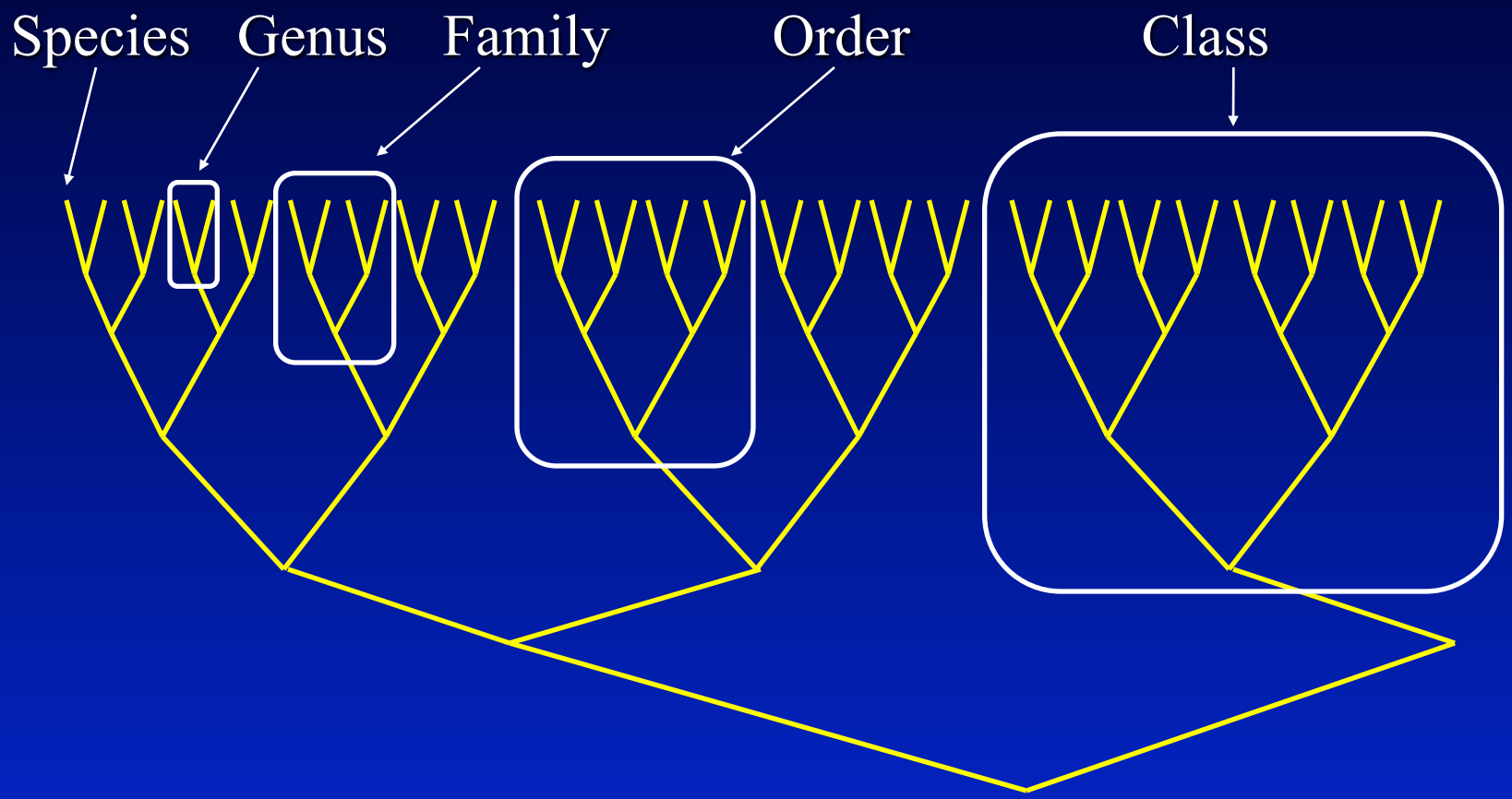

CAROLI LINNÆI
Naturæ Curioſorum *Dioſcoridis Secundi*
SYSTEMA
NATURÆ
IN QUO
NATURÆ REGNA TRIA,
SECUNDUM
CLASSES, ORDINES, GENERA, SPECIES,
SYSTEMATICE PROPONUNTUR.
Editio Secunda, Auctior.
STOCKHOLMIÆ
Apud GOTTFR. KIESEWETTER.
1740.

- Hierarchical system
  - Kingdom
  - Phylum
  - Class
  - Order
  - Family
  - Genus
  - Species

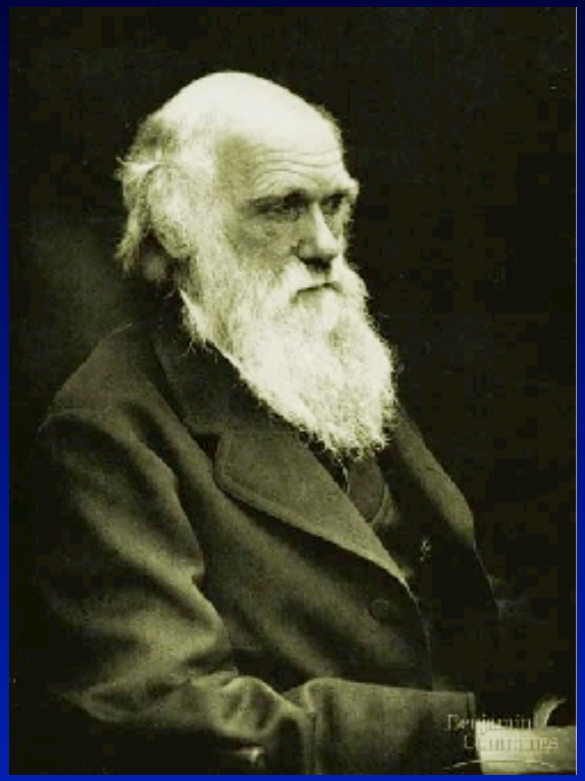
Classification depicted as a tree



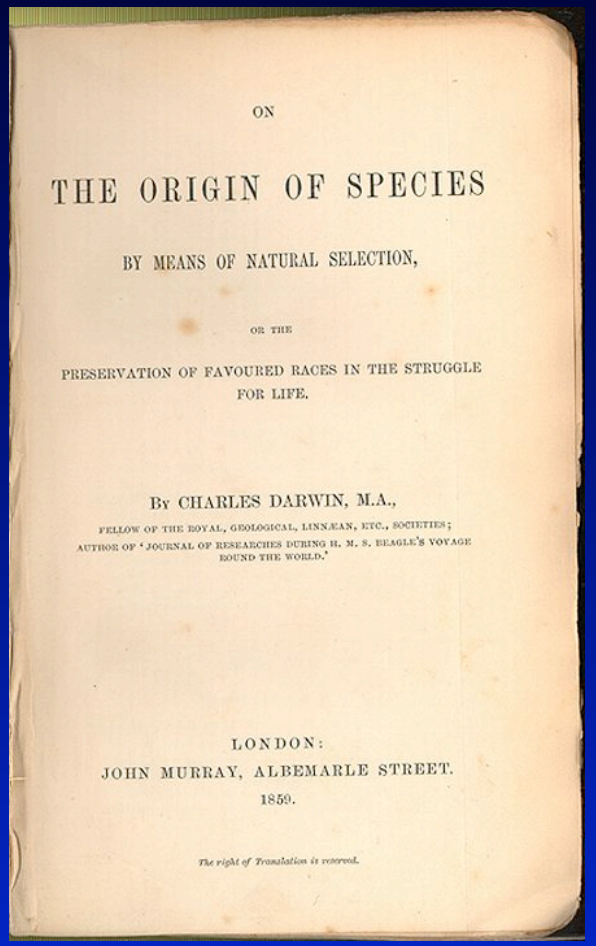
Classification depicted as a tree



Theory of evolution



Charles Darwin
1809-1882



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

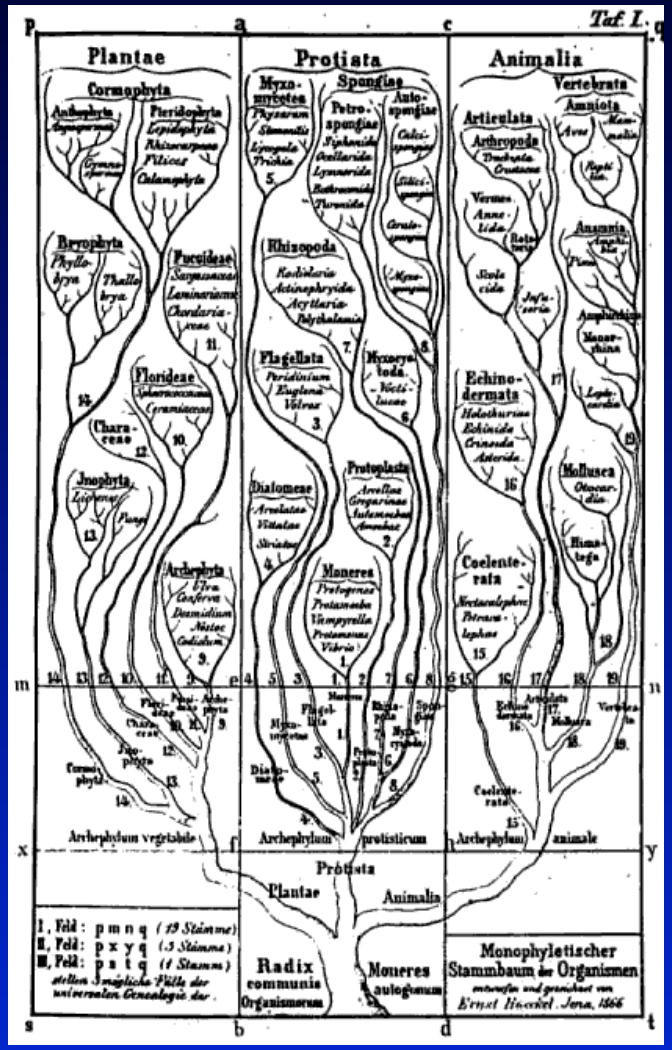
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

Phylogenetic basis of systematics

- Linnaeus:
  Ordering principle is God.
- Darwin:
  Ordering principle is shared descent from common ancestors.
- Today, systematics is explicitly based on phylogeny.



Natural Selection: Darwin's four postulates

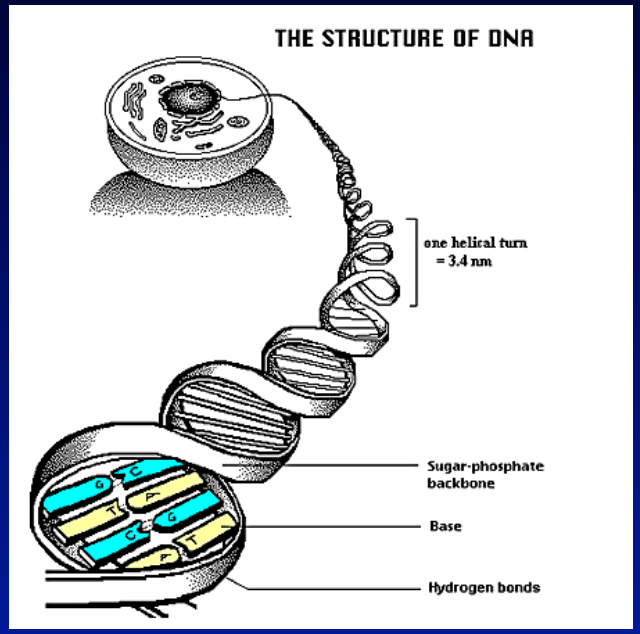
- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.

- Evolution by means of natural selection
- Presence of "design-like" features in organisms:
- Quite often features are there "for a reason"

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

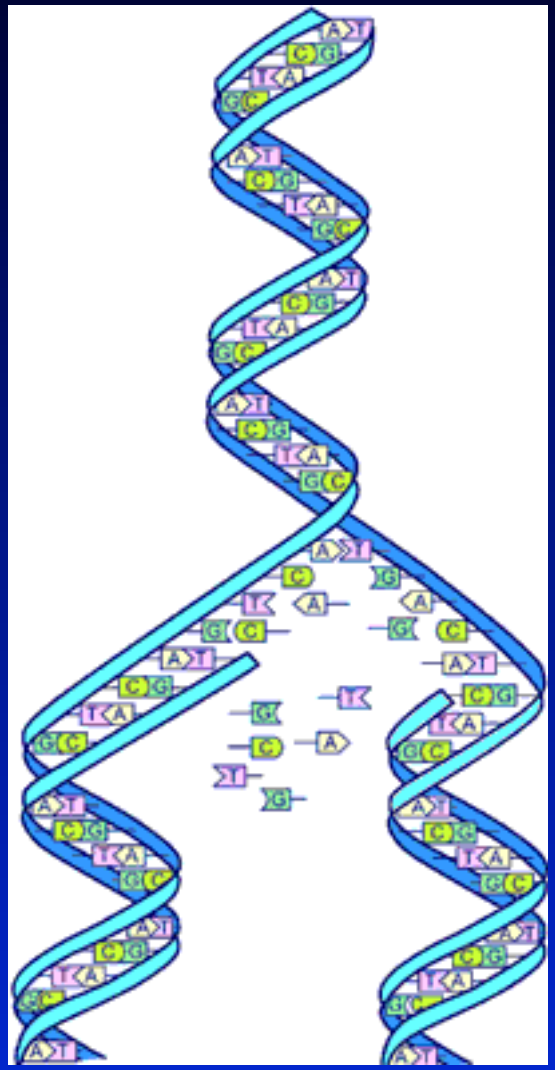
Bioinformatics

What is bioinformatics?



Bioinformatics: computational analysis of biological data

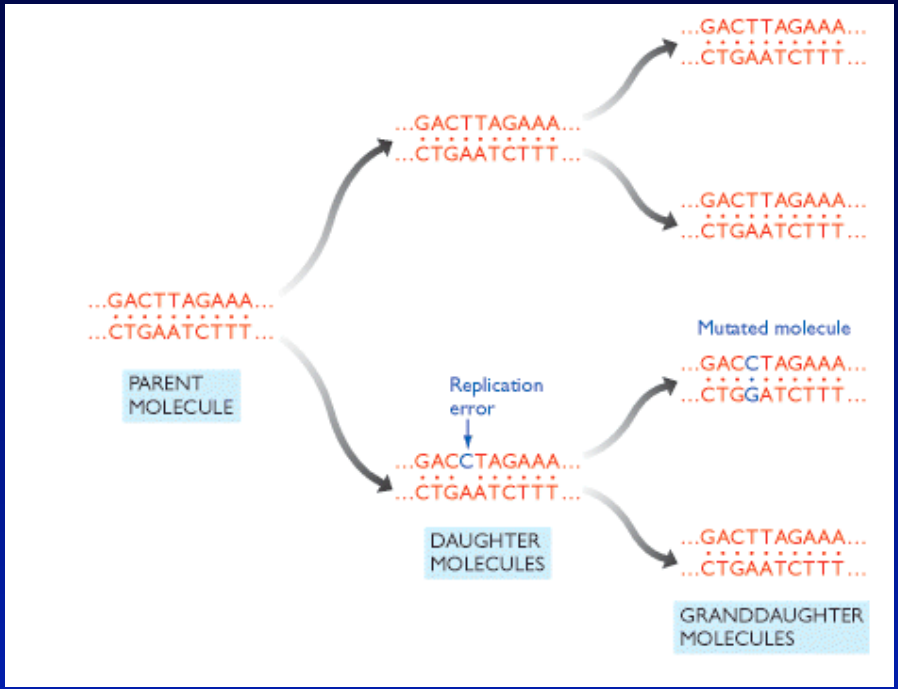
Molecular Basis for Heredity: DNA



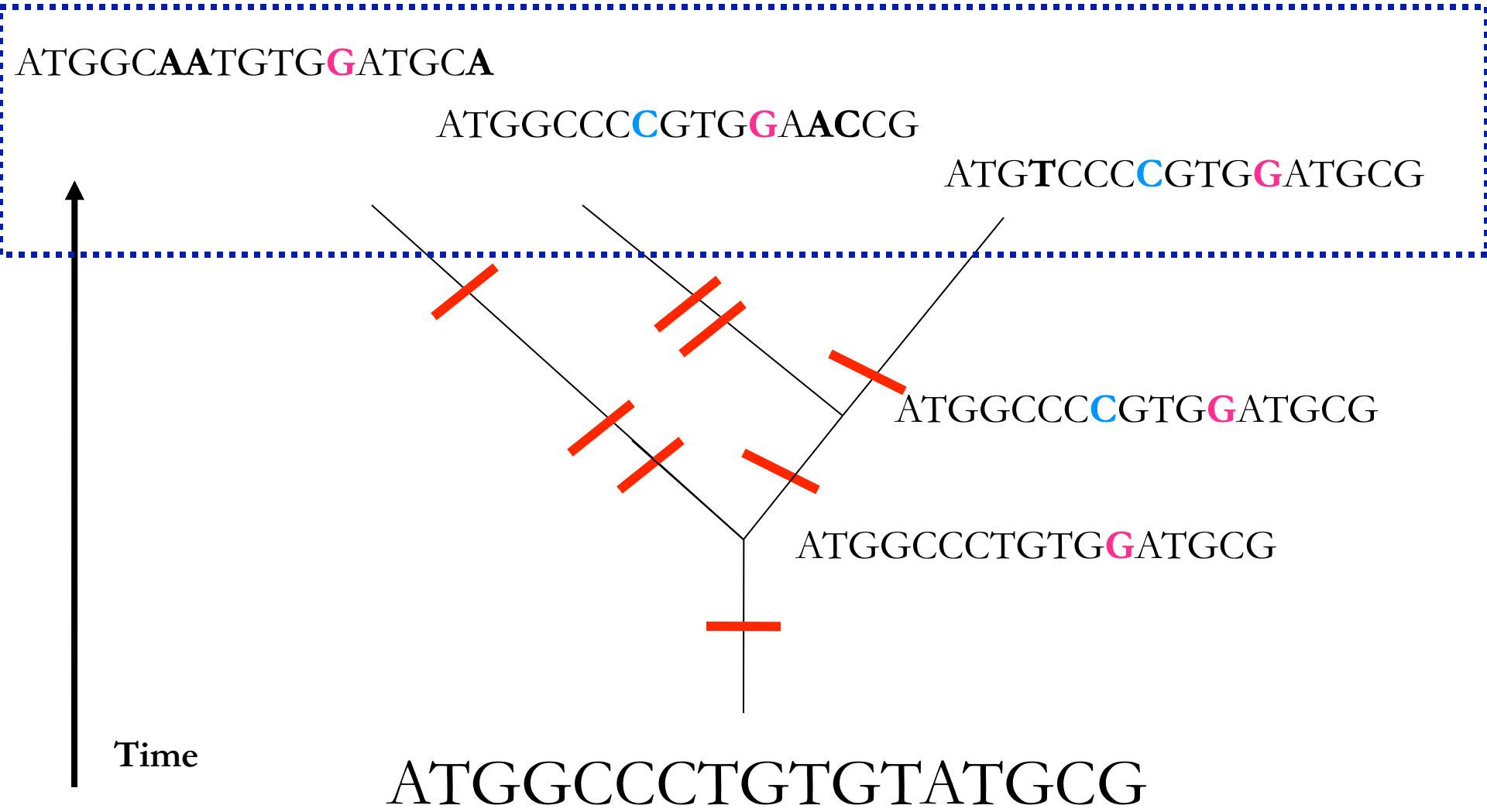
A always paired with T

C always paired with G

Molecular Basis for Evolution: DNA Mutation



A history of mutations

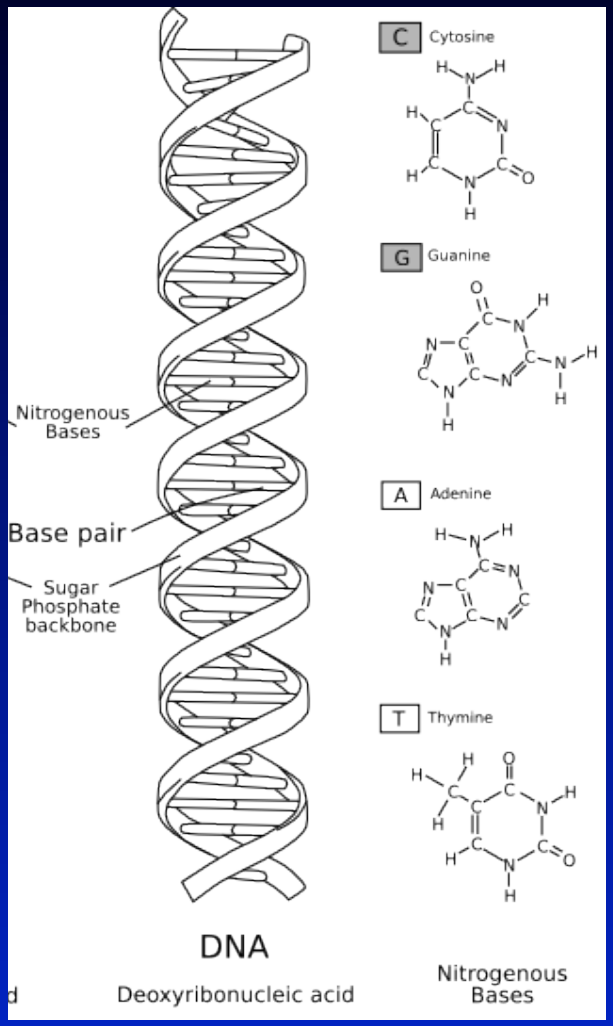


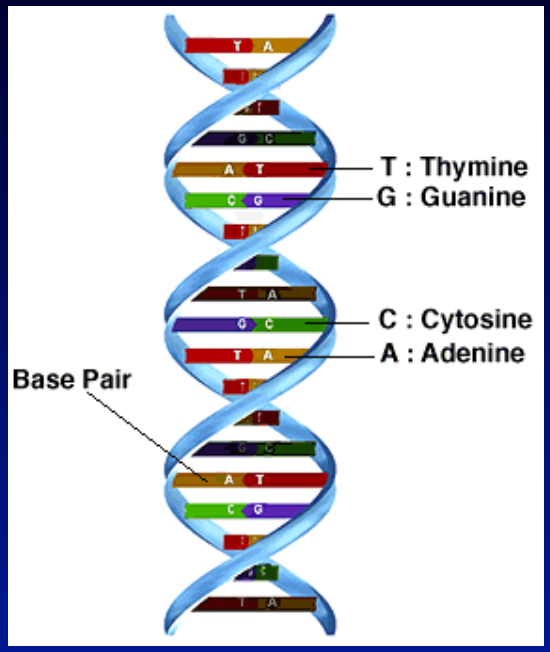
“DNA alignment”

- Species1: ATGGC**AA**TGTG**G**ATGC**A**
- Species2: ATGGCCC**C**GTG**G**A**AC**CG
- Species3: ATG**T**CCC**C**GTG**G**ATGCG

6
3
5

Symbolic representation of DNA structure

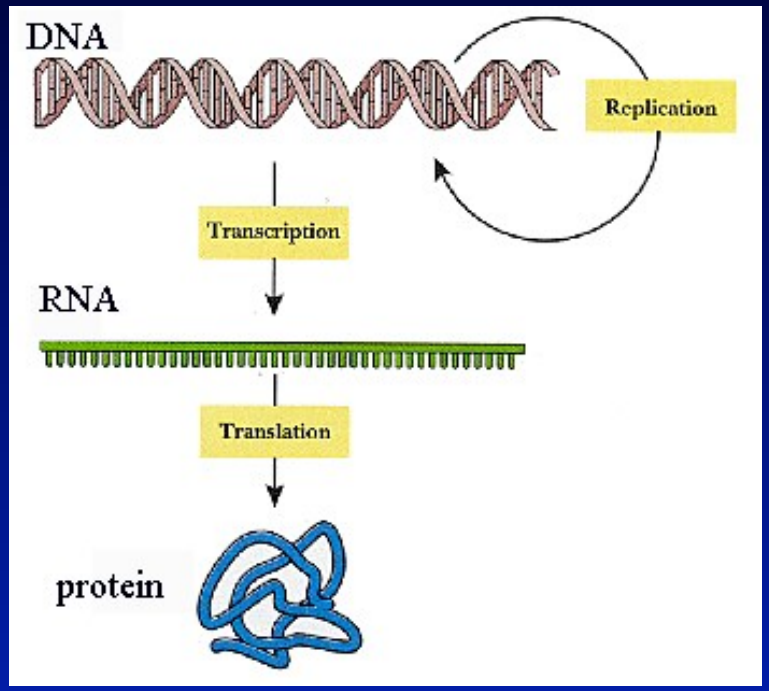




- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/ linguistically

HIV genome (approximately 10.000 nucleotides)

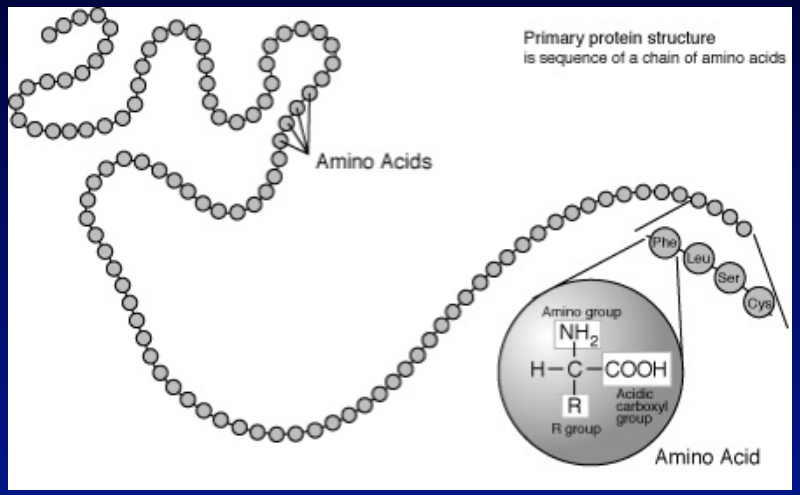
DNA --> RNA --> protein

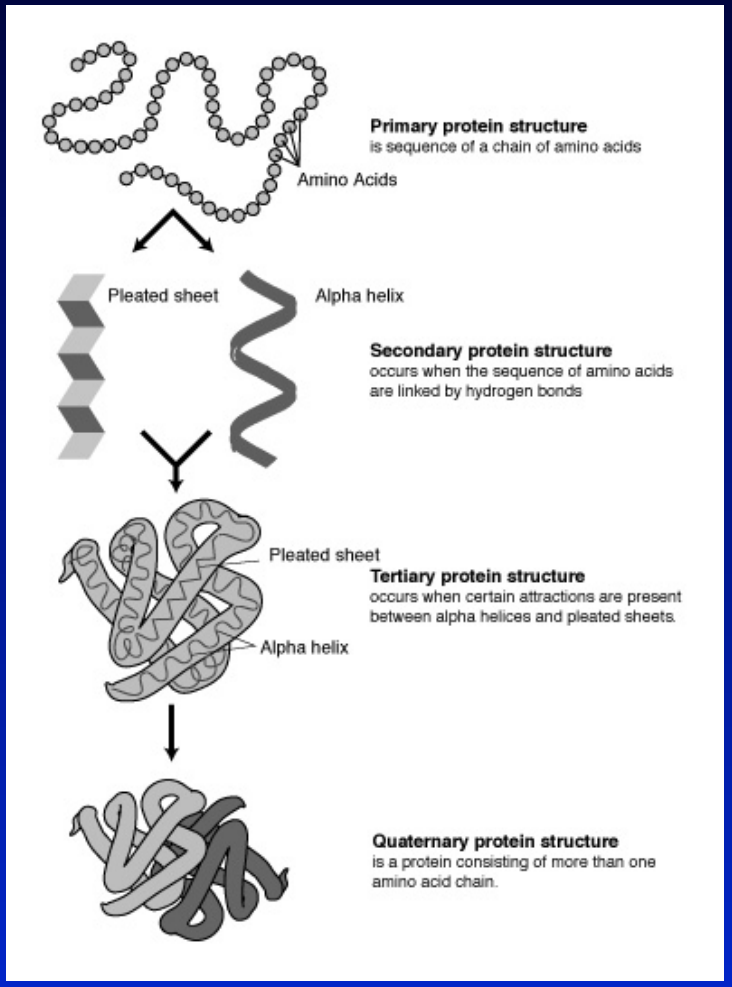


Standard Genetic Code

	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och	*	TGA	Opa	*	A
	TTG	*Leu*	*L*	TCG	Ser	S	TAG	Amb	*	TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	*Leu*	*L*	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	*Met*	*M*	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure

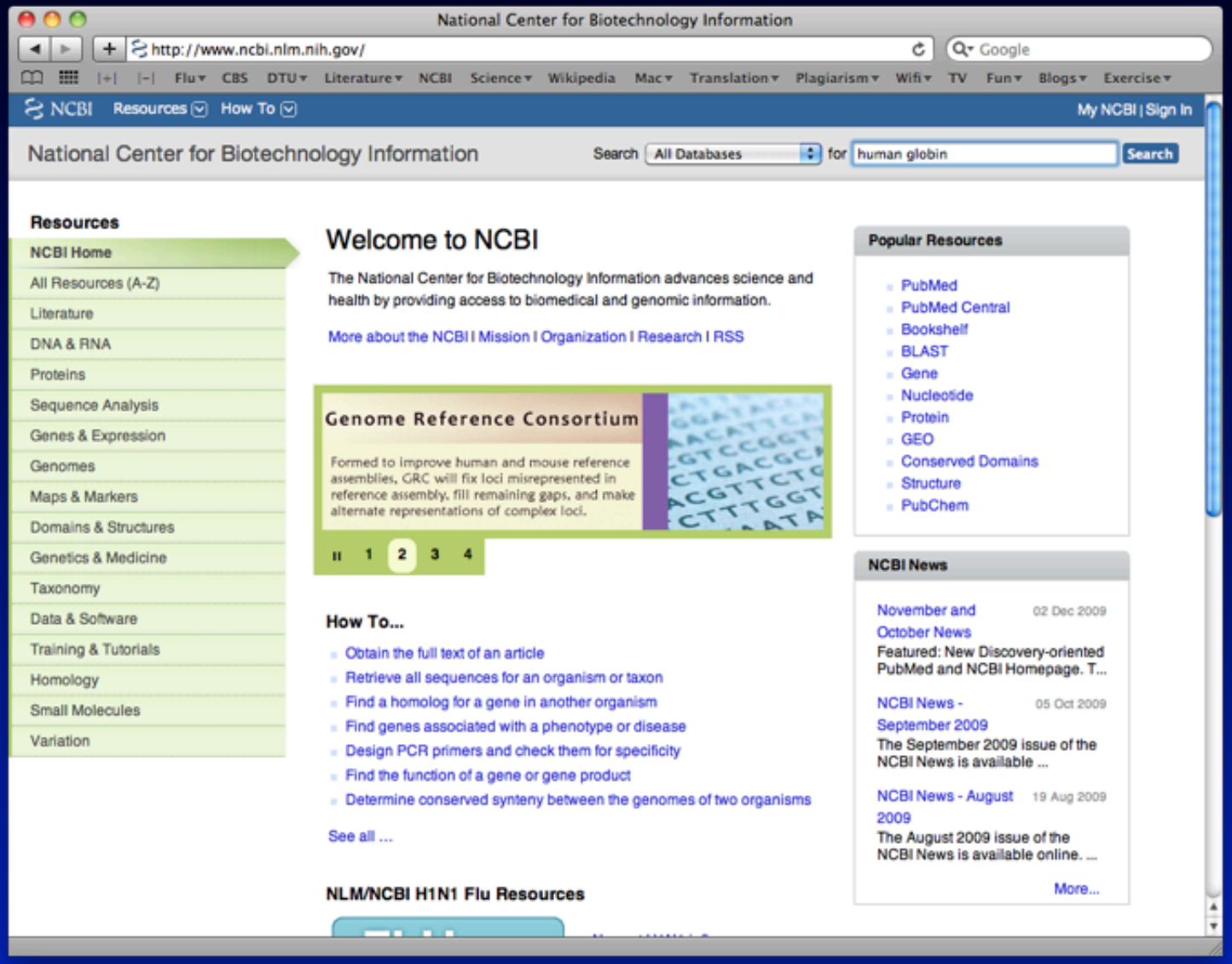




- Proteins are linear polymers
- Built from 20 amino acids
- Can be represented as string of 20 symbols

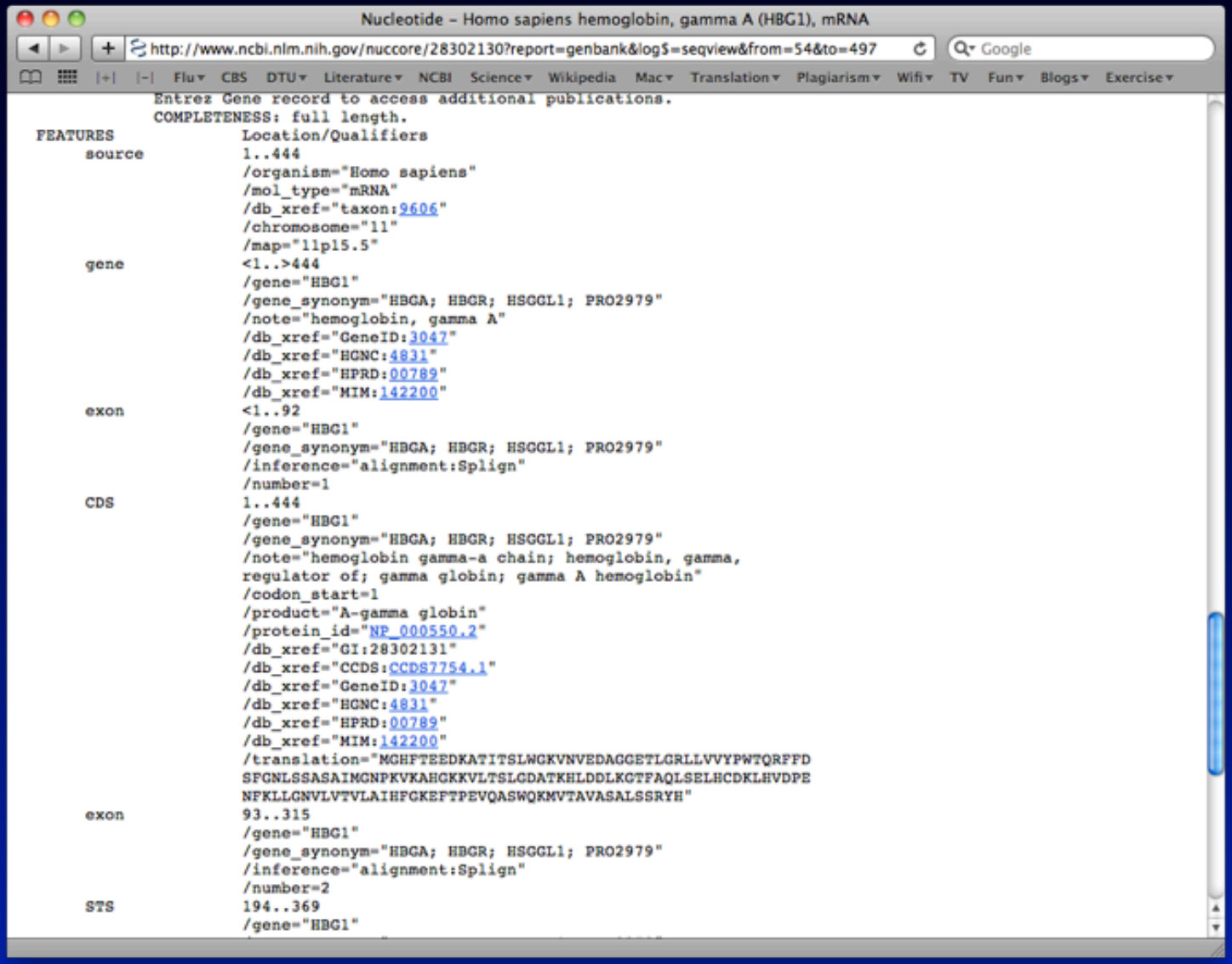
  ACDEFGHIKLMNPQRSTVWY

NCBI databases

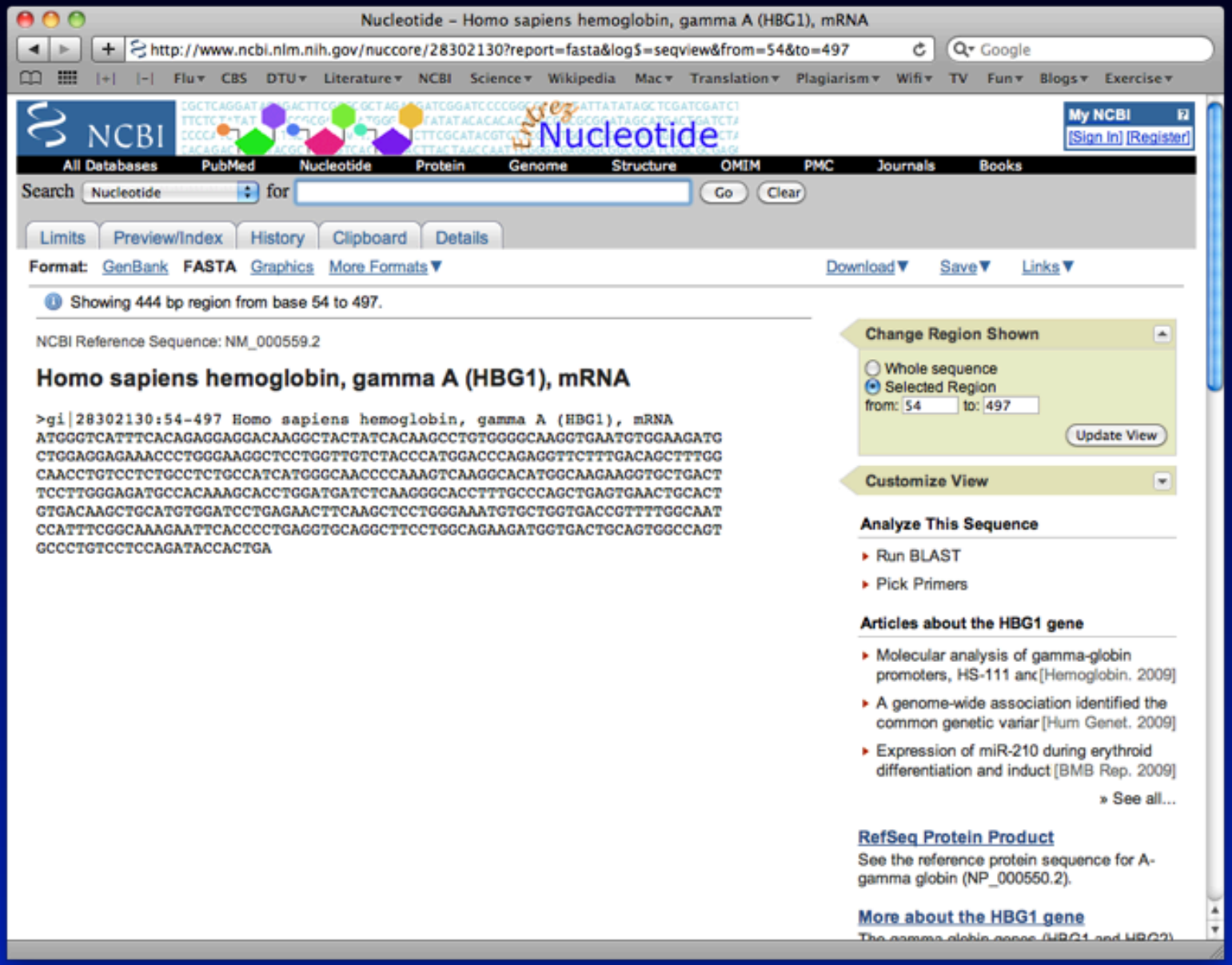


CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

NCBI databases: Genbank feature table



NCBI databases: fasta format



FASTA file